PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20244
doi: 10.3897/tdwgproceedings.1.20244

OPEN ACCESS

Conference Abstract

Integrating data-cleaning with data analysis to enhance usability of biodiversity big-data

Tomer Gueta‡, Yohay Carmel‡

‡ The Technion – Israel Institute of Technology, Department of Civil and Environmental Engineering, Haifa, Israel

Corresponding author: Tomer Gueta (tomer.gu@gmail.com)

Received: 13 Aug 2017 | Published: 14 Aug 2017

Citation: Gueta T, Carmel Y (2017) Integrating data-cleaning with data analysis to enhance usability of biodiversity big-data. Proceedings of TDWG 1: e20244. https://doi.org/10.3897/tdwgproceedings.1.20244

Abstract

Biodiversity big-data (BBD) has the potential to provide answers to some unresolved questions – at spatial and taxonomic swathes that were previously inaccessible. However, BBDs contain serious error and bias. Therefore, any study that uses BBD should ask whether data quality is sufficient to provide a reliable answer to the research question. We propose that the question of data quality and the research question could be addressed simultaneously, by binding data-cleaning to data analysis. The change in signal between the pre- and post-cleaning phases, in addition to the signal itself, can be used to evaluate the findings, their implications, and their robustness. This approach includes five steps:

1. Downloading raw occurrence data from a BBD.
2. Data analysis, statistical and / or simulation modeling in order to answer the research question, using the raw data after the necessary basic cleaning. This part is similar to the common practice.
3. Comprehensive data-cleaning.
4. Repeated data analysis using the cleaned data.
5. Comparing the results of steps 2 and 4 (i.e., before- and after data-cleaning). This comparison will address the issue of data quality, as well as answer the research question itself.

The results of step 2 alone may be misleading, due to the error and bias in the data. Even the results of step 4 may not be trustworthy, since data-cleaning is never complete, and

some of the error and much bias remain in the data. However, the changes in the results before- and after cleaning are important keys to answer the research question. If cleaned data reveal a stronger and clearer signal than raw data, then the signal is most likely trustworthy, and the respective hypothesis is confirmed. Conversely, if the cleaned data show a weaker signal than obtained from the raw data, then the respective hypothesis, even if confirmed by original data, needs to be rejected. Lastly, if there is a mixed trend, whereby in some cases the signal is stronger and in others it is weaker – the data is probably inadequate and findings cannot be considered conclusive. Thus, we propose that data-cleaning and data analysis should be conducted jointly.

We present a case study on the effects of environmental factors on species distribution, using GBIF data of all Australian mammals.

We used the performance of a species distribution model (SDM) as a proxy for the strength of environmental factors in determining gradients of species richness. We implemented three different SDM algorithms for 190 species in several different grid cells, that vary in their species richness. We examined the correlations between species richness and 10 different SDM performance indices. Species-environment affinity was weaker in species-rich areas, across all SDM algorithms. The results support the notion that the impact of environmental factors on species distribution at a continental scale decreases with increasing species richness. Seemingly, the results also support the continuum hypothesis, namely that in species-poor areas, species have strong affinities to particular niches, but this structure breaks in species-rich communities. Furthermore, a much stronger signal was revealed after data-cleaning. Thus, a joint study of a research question and data-cleaning provides a more reliable means for using BBDs.

Keywords

Data quality, New concept, GBIF data, Australian mammals, SDM

Presenting author

Tomer Gueta

Grant title

This research is supported by the Israel Science Foundation (ISF)